ICRA2026 Late Breaking Results Poster presented at
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Reinforcement Learning for Stair Locomotion of a Wheeled Bipedal Robot with Contact-Guided Behavior Cloning

Yi Gyeom Kim[1], Sejik Oh[1], Hyojin Jo[1], Dogyun Park[1] and Nam Kyu Kwon[2]
[1] Department of Robot AI Convergence, Yeungnam University, Gyeongsan 38541, Republic of Korea; dlrua01@yu.ac.kr, sjo7565@yu.ac.kr, hynn0254@yu.ac.kr, parkdk4634@yu.ac.kr
[2] Department of Electronic Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea; namkyu@yu.ac.kr

I. Introduction

This paper proposes a contact event-guided PPO with Behavior Cloning (PPO-BC) framework for stair locomotion of a 2-wheel 2-leg (2W2L) wheeled bipedal robot. Stair traversal is difficult because successful climbing depends on brief and sparse wheel-stair contact events that require precise leg lifting and posture stabilization. To address this issue, the proposed method trains a student policy using a combined objective of PPO-based reinforcement learning and behavior cloning from a pretrained frozen teacher policy. The teacher learns leg-centered climbing behaviors, while the student learns full 8-DoF control. A soft contact gate detects stair interaction directly from wheel contact forces and increases the BC contribution during critical contact phases without external terrain sensors. The method is validated under a minimal reward structure based on velocity tracking and postural stability, without stair-specific shaping rewards. Experiments in Isaac Lab simulation show that the proposed method outperforms both pure PPO and uniform PPO-BC in stair-crossing performance while maintaining stable locomotion after traversal.

II. Method

1. Teacher Policy Learning

To obtain an expert policy specialized for legged stair locomotion, a teacher policy was first trained using PPO. The original robot has an 8-DoF action space consisting of wheel and leg joint actions. During teacher training, the two wheel DoFs were disabled so that the policy learned only leg-centered locomotion behaviors. As a result, the teacher policy was forced to acquire stair-climbing motions using the six leg joint actions without relying on wheel actuation. This design allowed the teacher to learn expert behaviors for leg lifting, posture stabilization, and coordinated body motion during stair traversal. After training, the teacher policy was frozen and used as a reference policy for subsequent student learning.

2. PPO-BC Framework for Student Policy Learning

The overall framework of the proposed PPO-BC student learning process is shown in Figure. 1. A student policy was then trained to learn full 8-DoF control of the wheeled bipedal robot. Unlike the teacher, the student outputs both leg joint actions and wheel actions, enabling whole-body control for stair locomotion.

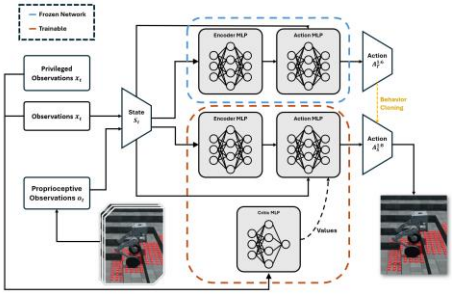

Figure1. Overall framework of PPO-BC

The student policy was optimized using a combined objective of PPO-based reinforcement learning and behavior cloning (BC) from the pretrained teacher policy. The PPO objective encouraged the student to maximize environmental rewards, while the BC objective guided the student toward the teacher's expert leg-centered behaviors.

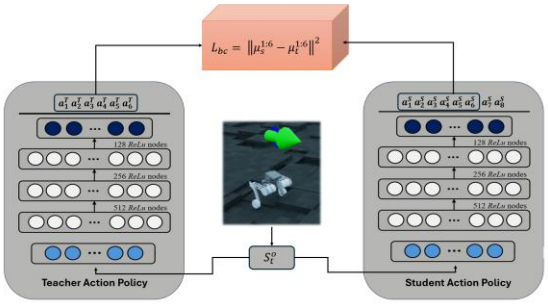

Figure2. BC loss computation in the PPO-BC framework.

As illustrated in Figure. 2, the BC loss was computed by minimizing the difference between the teacher's leg action outputs and the corresponding leg action outputs of the student policy. Since the teacher policy was trained as a leg-centered expert, the BC loss was applied only to the first six action dimensions corresponding to the leg joints. The BC loss is defined as

This research was supported by the Regional Innovation System & Education(RISE) program through the Gyeongbuk RISE CENTER, funded by the Ministry of Education(MOE) and the Gyeongsangbuk-do, Republic of Korea.(2026-RISE-15-115) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. RS-2023-00219725).

$$L_{bc} = \|\mu_s^{1:6} - \mu_t^{1:6}\|^2 \quad [1]$$

The PPO loss is defined as

$$L_{ppo} = L_{surrogate} + c_v L_{value} \quad [2]$$

The final objective for student training is given by

$$L_{total} = \lambda_{rl} L_{ppo} + \lambda_{bc} L_{bc} \quad [3]$$

where λ_{rl} and λ_{bc} are the weighting factors for reinforcement learning and behavior cloning, respectively.

3. Contact Event-Guided BC Weight Modulation

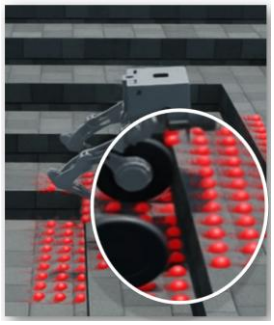

Figure3. Contact-trigger activation during wheel-stair interaction.

In stair locomotion, critical control behaviors are mainly required during wheel-stair interaction phases, while flat-terrain motion can be handled effectively through standard PPO-based optimization. To exploit this property, the proposed method introduced a contact-triggered modulation mechanism that dynamically adjusted the BC contribution according to contact events. As illustrated in Figure. 3, a contact-trigger is activated when the wheel interacts with the stair edge during climbing.

The upward contact gate was defined as

$$gate_{up} = clamp\left(\frac{f_{xy} - \tau_{low}^{up}}{\tau_{high}^{up} - \tau_{low}^{up}}, 0{,}1\right) \quad [4]$$

and the downward contact gate was defined as

$$gate_{down} = clamp\left(\frac{f_z - \tau_{low}^{down}}{\tau_{high}^{down} - \tau_{low}^{down}}, 0{,}1\right) \quad [5]$$

where f_{xy} denotes the horizontal wheel contact response and f_z denotes the vertical wheel contact response. Based on these signals, the BC weight was dynamically modulated as

$$w = \left(1 + gate_{up}\right) \cdot clamp(1 - gate_{down},\ 0.2,\ 1) \quad [6]$$

Using the modulated weight, the final training objective was defined as

$$L_{total} = \lambda_{ppo} \cdot L_{ppo} + \lambda_{bc} \cdot w \cdot L_{bc} \quad [7]$$

where λ_{ppo} and λ_{bc} are the weighting factors for PPO and behavior cloning, respectively. This formulation increased imitation guidance during critical stair-contact phases while reducing unnecessary imitation pressure during non-contact phases. As a result, the student policy received stronger supervision only when precise leg control was required, without relying on external terrain sensors or stair-specific shaping rewards.

III. Results

Table 1. Success rates (%) at different stair heights.

	5cm	10cm	15cm
PPO	99.22%	73.08%	11.22%
PPO-BC	100%	96.58%	74.23%
Proposed	100%	97.66%	91.01%

Training was performed with 4096 parallel environments. The teacher policy was trained for 15,000 iterations, followed by 15,000 iterations of student training using the PPO-BC framework. The simulation environment included both flat-ground and stair-traversal phases so that the policy could learn both normal driving behavior and contact-sensitive stair locomotion within a unified setting.

To evaluate the effectiveness of the proposed method, three learning configurations were compared: pure PPO, uniform PPO-BC, and the proposed contact event-guided PPO-BC. Pure PPO used only reinforcement learning without teacher supervision. PPO-BC used both PPO and behavior cloning with a fixed BC weight over all time steps. In contrast, the proposed method dynamically modulated the BC contribution according to wheel contact events.

Performance was evaluated at stair heights of 5 cm, 10 cm, and 15 cm. A trial was regarded as successful when the robot climbed the stair and maintained stable locomotion without falling during the remaining episode. Through this setup, the experiments focused on whether the proposed method could improve stair-climbing success under a minimal reward structure without stair-specific shaping rewards.

IV. Conclusion

This work proposed a contact event-guided PPO-BC framework for stair locomotion of a wheeled bipedal robot and demonstrated improved stair-climbing success over PPO and uniform PPO-BC in simulation. However, the current validation was limited to simulation-based success-rate comparisons. Future work will include more diverse terrain settings and real-robot experiments to evaluate robustness and practical applicability..